A method for grouping synonyms

Ingrid Falk	Claire Gardent	Evelyne Jacquey	Fabienne Venant
INRIA &	CNRS &	CNRS &	Université Nancy 2 &
Université Nancy 2	INRIA Nancy Grand-Est	ATILF	INRIA Nancy Grand-Est
falk@loria.fr	gardent@loria.fr	jacquey@atilf.fr	venant@loria.fr

Introduction

Because the Princeton WordNet has proved a valuable resource in NLP, many approaches have been developed to support the automatic creation of WordNets for languages other than English. In this paper, we present a method for grouping synonyms and definitions which we believe, can provide the basis for a *merge* approach to WordNet creation, that is an approach which starts by defining synsets (groups of synonyms) and then structures them into a WordNet type network. More specifically, given a word *w*, a set Syn_w of synonyms for *w* and a set of definitions $\{d_1, \ldots, d_n\}$ representing the possible meanings of *w*, the proposed method associates subsets of Syn_w with definitions of *w*. That is, it permits dividing a set of synonyms into subsets and associating each of these subsets with a definition.

Method

Gloss based similarity measures have been used for word sense disambiguation (WSD) [PBP05] to associate a word occurring in a given linguistic context to a given meaning. Similarly, we apply gloss based similarity measures to disambiguate synonyms that is, to choose for each synonym Syn_i of a word *w*, the meaning (represented by a definition d_i of *w)* which fits best.

The definitions used are extracted from a general purpose dictionary for French called the TLFI (Trésor de la Langue Française informatisé). Similarly, the synonyms of a given word *w* are extracted from a synonym dictionary for French called Le Petit Robert. For this experiment, we focused on verbs.

To divide the set of synonyms associated by Le Petit Robert synonym dictionary into subsets and associate each resulting synonym subset with a TLFI definition, we proceed as follows:

Index creation. We derived from each TLFI definition an index consisting of the set of lemmatised open class words occurring in the definition.

Computing similarity scores. Given a verb *v*, Syn_v the set of its Petit Robert synonyms and D_v, the set of its TLFI definitions we compute for each pair $<d_i , syn_j>$ such that $d_i \in D_v$ and $syn_j \in Syn_v$ a similarity score using gloss based similarity measures.

Synonym/Definition matching. For each synonym syn_j of a verb *v* with definitions D_v, we associate syn_j with the definition $d_i \in D_v$ for which the similarity measures give the highest (non null) score.

To assess the impact of the similarity method used, we applied two kinds of similarity measures: word overlap and vector based measures. Word overlap based measures were introduced by [Les86] to perform word sense disambiguation. For a given verb, we compare the index of each of its definitions with the merged indexes of a synonyms definitions. That definition which has the most words in common with the synonyms definitions is chosen as its most appropriate sense. Vector based similarity measures represent a text as a vector in a word space. Then a TLFI definition can be seen as representing a “direction” in the word space and

similarity between words can be computed using some vector similarity measure. The overlap measures we use are *simple word overlap*, *extended word overlap* and *extended word overlap normalised*. The vector based measures used are *local word vectors*, and *second order word vectors* with and without a tf*idf cutoff [PBP05].

Evaluation.

To evaluate the groupings produced by our method, we built a gold standard consisting of 27 verbs differing in their position (high, medium, low) on three scales (polysemy, genericity and frequency). For each verb, the association between definitions and synonyms was done by 4 professional lexicographers with an interannotator agreement rate of 87%. The resulting reference associates then for each verb in the chosen sample, the set of TLFI definitions and with each definition, a set of synonyms.

To compute recall and precision, we extract the set of triples $<v, syn_i, def_j>$ defined by the reference such that syn_i is a synonym of v which has been assigned to definition def_j by the annotators. Recall is then the number of correct tuples produced by the system divided by the total number of reference tuples and precision is the same number divided by the total number of tuples produced by the system. F-measure is the harmonic mean of precision and recall. The baseline gives the results obtained when randomly assigning the synonyms of a verb to its definitions.

Results

The 6 similarity measures tried all performed similarly with recall values neighbouring 0.7 (baseline 0.43) and precision 0.75 (baseline 0.44). The best F-measure score (0.71) was achieved with a word overlap measure (*extended word overlap normalised)*.

These results indicate that the TLFI definitions are sufficiently rich to support the application of gloss based similarity measures to the task at hand. That is, even though definitions are typically short, when considering the definitions of two synonyms, the word overlap between two TLFI definitions is sufficiently large to support a meaningful division of the set of synonyms for a given word *w* into subsets of synonyms corresponding to the several possible meanings of that word.

Conclusion

The method described here provides a principled way of constructing out of a set of synonyms, a set of synonym subsets each labeled with a (TLFI) definition. Furthermore, it has two features which we believe, make it appropriate as a basis for WordNet construction.

First, it differs from much work on automatic synonym extraction or WordNet construction in that it avoids introducing noise in the data and does not group together words that are not synonyms. This is because it takes as a basis synonym dictionaries. Second, it can be used to merge the content of several synonym dictionaries in a meaningful way. This is because synonym grouping is here mediated by a synonym-to-definition mapping which is independent of the particular synonym groupings listed by synonym lexicons. Indeed, we are currently applying the method described here to merge five synonym lexicons for French. Further extension of the coverage can be achieved by applying the same method to any available synonym dictionary such as for instance, the Wiktionary[1] .

Once sufficient coverage is achieved, the question arises of how the synonym subsets obtained can be linked in a WordNet like structure. We plan to investigate two possible

1 http://fr.wiktionary.org/wiki/

directions. One possibility is to use ontology fusion methods for merging our synsets with either the French EuroWordNet or WOLF [SF08, FS08]. Another possibility consists in combining our method with a translation approach in order to associate each <verb, definition> pair to a Princeton synset. In this way, we can build on the WordNet structure given by PWN and enrich the synsets derived from the five synonym dictionaries with translations of the related english synonyms.

References

[FS08] D. Fiser and B. Sagot. Combining multiple resources to build reliable wordnets. In *Proc. of TSD*, Brno, Tchéquie, 2008.

[Les86] M. Lesk. Word sense disambiguation: Algorithms and applications. In *Proceedings of SIGDOC*, 1986.

[PBP05] T. Pedersen, S. Banerjee, and S. Patwardhan. Maximizing semantic relatedness to perform word sense disambiguation. Technical report, University of Minnesota, 2005.

[SF08] B. Sagot and D. Fiser. Building a free french wordnet from multilingual resources. In *Proc. of Ontolex*, Marrakech, Maroc, 2008.